Towards Evaluation of Inferred Gene Network

S. Zainudin
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia
Email: suhaila.zainudin@gmail.com

S.Deris
Sekolah Pengajian Siswazah
Universiti Teknologi Malaysia, UTM Skudai 81310,
Johor, Malaysia

Abstract

Gene network is a representation for gene interactions. A gene collaborates with other genes in order to function. Past researches have successfully inferred gene network from gene expression microarray data. Gene expression microarray data represent different levels of gene expressions for organisms during biological activity such as cell cycle. A framework for gene network inference is to normalize gene expression data, discretize data, learn gene network and evaluate gene interactions. This framework was used to learn the gene network for two S.cerevisiae gene expression datasets (Spellman Cell cycle and Gasch Yeast Stress). Gene interaction inference was also done on data contained in 8 major clusters found by Spellman. The inferred networks were compared to gene interaction data curated by Biogrid. Results from the comparison shows that some of the inferred gene interactions agree with data contained in Biogrid and by referring to curated genetic interactions in Biogrid, we can understand the significance of computationally inferred gene interactions.

Keywords: Bayesian Network, Microarray Gene Expression, Gene Interactions, Gene Network.

1. Introduction

System Biology is an emerging branch of research committed to studying biological systems which is a set of biological objects (proteins and genes) involved in a biological process. A technique often used in System Biology is graphical model. Graphical model is commonly used to represent qualitative models of biological systems.

A biological system is described in the form of graphs using graphical models. A graph is useful for representing dependencies in a biological system such as gene A activates gene B or gene C and gene D suppress gene E. An example of graphical model is Bayesian Network which can encode probabilistic dependencies among a set of variables. We can use Bayesian Network to simultaneously study many genes which are active during experiments such as cell cycle or responding to external stress.

Traditionally, earlier methods in Molecular Biology focused on studying a single gene in a single experiment. Although this method enabled us to fully understand a gene, it does not illustrate how a gene will function in collaboration with other genes since a gene do not work in isolation. Commonly, genes collaborate with other genes to carry out biological processes in an organism. Even a simple organism such as yeast contains over 6000 genes. Hence to completely understand gene interactions amongst this large number of genes, we need a high-throughput technique to simultaneously analyze thousands of gene expressions. This will help us understand activities of whole genome, instead of only understanding a single or a few genes as was done previously.

Graphical models provide a top down picture of gene interactions which provides a better view of the functional role of genes. We can use graphical models to answer one of the most important question in biology; how gene expression is turned on or off. Most cells in an organism have an identical genome. However, different levels of gene expression are produced by each gene according to different needs.

Gene expression is the process where the information coded in genes is transcribed into mRNA and then translated into proteins. Microarray technique has been extensively used to capture the expressions of thousands of genes under various external stimuli. For example, Spellman conducted a microarray experiment to catalog yeast genes whose transcript vary periodically within the cell cycle [1]. The experiments captured the complex interactions between genes in the form of levels of gene expression at different time points. Available machine learning techniques are then used to learn gene interaction models in the form of a gene network [2][3][4].

0-7695-2945-3/07 $25.00 © 2007 IEEE
DOI 10.1109/ICCSA.2007.29

Gene network refers to a set of molecular components such as genes, proteins and other molecules, and interactions between them that collectively carry out some cellular function. Such networks are increasingly used as models to represent phenomenon at gene expression level [5]. Gene network provides an overview of the physiological state of an organism at mRNA level. It also provides a system view of gene activities. Other than that, gene network could be used to describe functions and as the means for annotation of genomics and functional genomics data. Gene network may also be used to uncover the complete biochemical networks of cells. Some of the techniques that were used to infer gene network from gene expression data are Boolean network, Bayesian Network and Differential Equation. A review on these techniques and more can be found in [6][7].

An effective technique often used to infer gene networks is Bayesian Network. A Bayesian Network (BN) is a graphical representation of a probability distribution. It is a compact and intuitive representation. From its application in other domains, it is useful for describing processes composed of locally interacting components. BN has a solid statistical foundation since it is based on the Bayes' Rule. BN also has an efficient model learning algorithm and can capture causal relationships. It is also able to deal with noisy data (a normal occurrence for microarray data). BN is suitable to the problem of gene network inference since gene expression is a stochastic(non-deterministic) phenomenon.

A BN is composed of two components; a Directed Acyclic Graph (DAG) and a Conditional Probability Distribution (CPD). A DAG contains nodes which represent random variables and edges which represent relation between genes. In our context, a DAG contains nodes which represent genes and edges which represent interaction between genes. The Conditional Probability Distribution(CPD) contains the set of parameters (gene expression values) for the DAG. Attempts to infer gene network using BN are some of the most successful so far. However, more accurate network inference can be achieved by combining BN with other technique or by using more than one type of data. In the early days, only microarray data was used to infer gene network, however, now most researchers include other types of data such as protein-protein interaction data [8] and promoter detection element[9]. Using multiple types of data has been shown to produce more accurate network in less amount of time [10].

Next, we shall discuss the evaluation of inferred gene network.

2. Inferred Results Evaluation Problem

Extensive work has been done in gene network inference. However, little work were done on evaluating inferred results [4]. According to [12], gene interactions inference problem is hard since we are trying to learn interactions amongst hundreds of genes from limited datasets which contain only several time points in an experiment. This problem involves statistical robustness since a small sample number is not enough to differentiate between true and spurious interactions [11]. Some solution to this problem is using methods such as bootstrap[16] to identify network features which can withstand perturbations to observations.

Another solution is using prior knowledge about biological principles to restrict the set of network structures to be considered. We can also restrict further by evaluating much smaller structure sets on the basis of prior knowledge about specific genes such combining genes with their respective regulators in a set.

Research to come up with a way to formulate firm proof of the significance of inferred gene network has not been done. Studies to assess inferred interactions commonly used own gold standard compilation. This leads to two main problems, estimation of sensitivity and assessment of false interactions. Researchers tend to use biological literature to back up their claims, such as using protein sequence similarity to imply similar gene functions[12]. Although sequence similarity does not necessarily do so. For the second problem, when we inferred an interaction which is not contained in the literature, we cannot decide without further costly gene knock out experiments whether the inferred interactions are spurious(false) or newly discovered unknown interactions [12].

3. Current Evaluation Approaches

Table 1 is a collection on some gene network evaluation approach. This review on evaluation methods has prompted the search for a reliable repository of gene interactions to help in understanding the gene interactions which has been successfully inferred using BN. We chose to use Biogrid (www.thebiogrid.org). This repository provide access to large datasets of biological interactions that are important to gene and protein studies especially in area of function study and analysis of global network properties. Datasets that are affiliated with Biogrid are SGD (www.yeastgenome.org), GeneDB (www.genedb.org), Flybase (flybase.bio.indiana.edu) ,WormBase (www.wormbase.org), GeneOntology (www.geneontology.org), Human Protein Reference Database (www.hprd.org) and others.

TABLE I
EVALUATION METHODS FOR INFERRED GENE NETWORK

RESEARCH	EVALUATION
Friedman 2004[11]	Systematic evaluation against literature
Ott 2004[4]	Compare to simulated gene expression data from artificial network
Husmeier 2003[12]	Use ROC graph to measure sensitivity and specificity of inferred edges.
Lee and Lee 2005[13]	Compare with literature
Kim et al 2003[14]	Compare with cell cycle pathway in KEGG

In the coming section, we will explain the process of inferring gene network.

4. Gene Network Inference Framework

Based on current literature, Bayesian Network is chosen as the representation for inferred gene network in this study. Subsequently, the framework for this study is in Figure 1. The data was downloaded from S.cerevisiae Genome Database (www.yeastgenome.org/). The data was pre-processed during which any missing data is imputed using Mean Completer from Rosetta (http://rosetta.lcb.uu.se/). This method substitutes missing values with the mean value for all observed expression values for a gene. Then data was discretized into 3 levels; Up (represents gene over-expression), Normal (represents unchanged gene expression) and Down (represents under-expression).

Learning utilizes PNL (Probabilistic Network Library) which was developed by Intel (http://www.intel.com/technology/computing/pnl/). The learning engine utilized pre-processed data to learn Bayesian Networks in the form of Directed Acyclic Graph (DAG) and Conditional Probability Distribution (CPD). The learning type employed is Maximum Likelihood. This learning type will search for the best model that best explains the data. Searching for best model uses hill-climbing approach. Hill-climbing starts with an initial DAG, then it make changes to initial DAG such as adding an arc, deleting an arc or reversing an arc to get a new DAG. Then, compare new DAG's Bayesian Information Criterion (BIC) score to initial DAG's BIC score. If new DAG's score is more than initial DAG's score, new DAG is returned as the best DAG. If new DAG's score is less than initial DAG's score, the initial DAG is retained as best DAG. The best DAG can be viewed using Graphviz (www.graphviz.org).

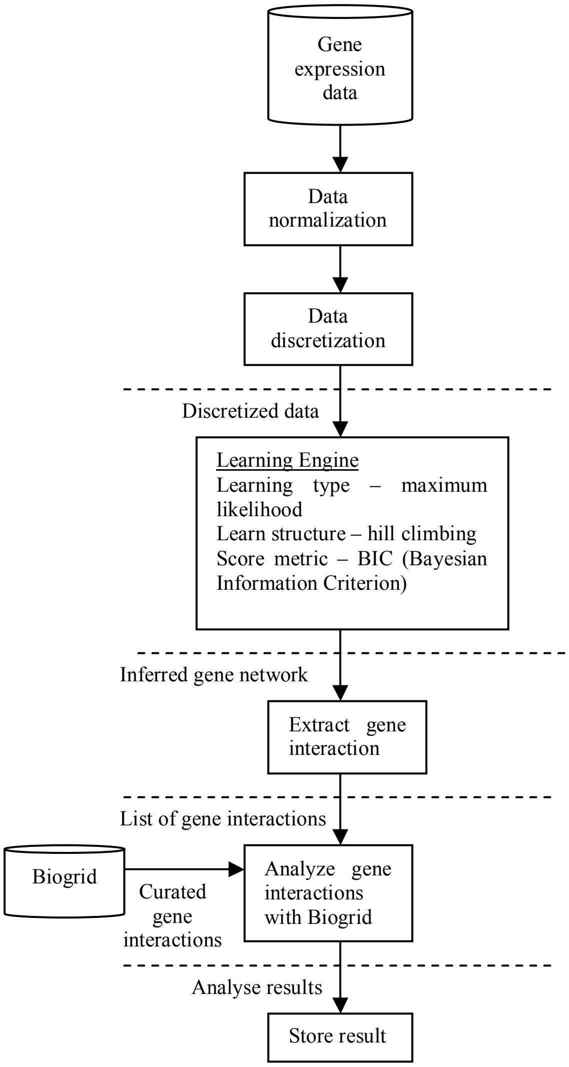
Gene expression data
Data normalization
Data discretization
Discretized data
Learning Engine
Learning type – maximum likelihood
Learn structure – hill climbing
Score metric – BIC (Bayesian Information Criterion)
Inferred gene network
Extract gene interaction
List of gene interactions
Biogrid
Curated gene interactions
Analyze gene interactions with Biogrid
Analyse results
Store result

Figure 1. Gene network inference framework

In order to analyze gene interactions contained in the network, a list of interactions is extracted from dot files for inferred gene network. Then, the list is compared to Biogrid curated gene interaction list to find any similar interactions between inferred gene interactions and curated gene interactions.

We have discussed gene network inference framework in general. In the next section, we shall concentrate on the details about learning gene network.

5. Learning the Gene Network

The strategy to learn the gene network is explained here. Network is learned in the form of Bayesian Network in which a DAG represents the network of genes from data and a CPD is calculated from the gene expression levels from data. The learning approach is below.

Step 1. Create the initial gene network(initial_network) with given initial CPD values (initial_CPD).
Step 2. Create a linked structure as the starting point for structure learning (linked_structure).
Step 3. Create a Bayesian network (b_net)with random matrices (using linked_structure and initial_CPD from initial_network).
Step 4. Create evidence_file
Step 5. Store evidence data from input file in evidence_file.
Step 6. Create learning engine using b_net as starting point.
Step 7. Run learning engine on evidence_file.
Step 8. Get the DAG (DAG_result).
Step 9. Get the DAG from initial_network (DAG_initial).
Step 10. Compare DAG_initial to initial_network. The DAG with best score will be returned as the best structure.

This approach was used to learn gene networks for Spellman Cell Cycle dataset (800 genes) and Gasch Yeast Stress dataset (38 genes). A total of 100 interactions are inferred from Spellman Cell Cycle dataset and 28 interactions were obtained from Yeast Stress dataset. The rest of the interactions learnt are included in Table 2. In the following section, we will give a brief overview of Biogrid which is used to evaluate the inferred interactions in this research.

6. Introduction to Biogrid

Biogrid is a freely accessible repository of protein and genetic interactions. Biogrid version 2.0 contains more than 116000 interactions from S.cerevisiae, C.elegans, D.melanogaster and H.sapiens. Over 30000 interactions were added from 5778 publications using detailed curations of the S.cerevisiae primary literature. Biogrid is based on Open software packages such as MySQL 4.1 for storing interactions and annotations from various resources, Fedora Core 3 as the operating system platform, PHP 5.0 and Apache 2.0 as the searchable front end and Python, Java and Perl which makes up the developers script and build tools.

7. Evaluation of Results

Inferred gene interaction were evaluated by comparing with 87922 curated gene interactions from Biogrid (version 2.0.22). The result of the comparison is in Table 2.

TABLE 2
COMPARISON RESULT FOR INFERRED GENE INTERACTION

Experiments	Total Interactions	Known Interactions	Unknown Interactions
800 genes	100	15	85
38 genes	28	2	26
CLB2 (35 genes)	22	3	19
Histone (9 genes)	8	4	4
MCM (37 genes)	22	2	20
SIC1 (26 genes)	21	1	20
CLN2 (56 genes)	22	3	19
MAT (13 genes)	10	0	10
MET(19 genes)	4	0	4
Y(31 genes)	22	0	22

In the following section, we shall discuss the significance of the inferred interactions from the biological point of view. The source for the discussion is from SGD and Biogrid. The discussion will focus on the interactions inferred from Spellman dataset only.

8. Inferred Interaction Details for 800 genes

Seven interactions were detected for this dataset which corresponds to interactions curated in Biogrid as shown in Table 3.

TABLE 3
DETAILS FOR INFERRED GENE INTERACTIONS FROM 800 GENES DATASET

Known Interactions	Occurrence	Authors
YBL002W->YBL003C	1	Krogan NJ et al.
YKL113C->YNL072W	2	Loeillet S et al. Pan X et al.
YLR103C->YPR135W	1	Tong AH et al.
YDR224C->YBR010W	1	Kurumizaka H et al.
YBR010W->YBR009C	4	Kurumizaka H et al. Sharp JA et al. Huang L et al. Sabet N et al.
YBR010W->YDR225W	1	Kurumizaka H et al.
YGR108W->YPR119W	3	Cross FR et al. Surana U et al. Richardson H et al. Honey S et al.

Interaction between YBL002W and YBL003C occurs because both are from histone subtypes. Histone is needed for chromatin assembly and chromosome function. Both genes has the same function which is DNA binding .

For interaction between YKL113C and YNL072W, the common link between both genes is both are active during Okazaki fragment processing and synthesis. Both genes are also involved in DNA replication process.

Regarding interaction between YLR103C and YPR135W, gene YLR103C is involved in the process by which DNA replication is started. Meanwhile gene YPR135W is involved with the process where new strands of DNA are synthesized. Hence it make sense that YLR103C interacts with YPR135W since after DNA replication is started, new DNA strands will be synthesized.

Interaction between genes YDR224C and YBR010W occurs because both are from histone subtypes required for chromatin assembly. Both genes has same DNA binding function.

Genes YBR010W and YBR009C are histone proteins which are required for chromatin assembly. Both share the same DNA binding function. Genes YBR010W and YDR225W are also from histone proteins involved in chromatin assembly and also carry out DNA binding function.

Both genes YGR108W and YPR119W shares the same function which is cyclin-dependent protein kinase regulator activity. Both are also B-type cyclin.

From interactions discussed above, most interactions occur for DNA binding function and DNA replication process.

9. Inferred Interaction Details for 28 genes

A total of 28 interactions were inferred from this dataset. From 28 interactions, 2 interactions were known and the other 26 were unknown interactions.

TABLE 4
DETAILS FOR INFERRED GENE INTERACTIONS FROM 38 GENES DATASET

KNOWN INTERACTIONS	OCCURRENCE	AUTHORS
YPL240C->YAL005C	1	Zhao R et al.
YPL240C->YBL075C	1	Marsh JA et al.

For genes YPL240C and YAL005C, both share the same molecular function, which is interacting selectively with an unfolded protein.

The interaction between genes YPL240C and YBL075C occurs because they share the same function which is unfolded protein binding. Both genes are also involved in the same process which is protein folding.

10. Conclusions and future directions

In this research, inferred gene interactions from microarray gene expression data have been successfully compared to Biogrid, a curated gene interactions repository. With the help from Biogrid which contains valuable information, we can understand computationally inferred interactions. Inferred gene interactions and curated gene interactions from wet lab experiments recorded in the literature are both vital in System Biology's quest to completely understand an organism's internal functions. We have inferred gene interactions using Bayesian Network and DNA microarray data. To obtain better results, we plan to use prior information and DNA microarray data together in future experiments. This study primarily used biological literature to evaluate inferred interactions.

The evaluation method used however cannot evaluate the performance of the algorithm itself in inferring true or false interactions. A possible method to evaluate the algorithm is by inferring a synthetic network using the structure learning algorithm. This approach can evaluate how much information from a known network can be recovered by the algorithm under different perturbations such as different number of sample and varying levels of noise [17] .

11. References

[1] P. T. Spellman, G. Sherlock, M. Q.Zhang, V. R. Iyer, K.Anders, M. B. Eisen, P. O'Brown, D. Botstein, and B.Futcher, "Comprehensive Identification of Cell Cycle-regulated Gene of the Yeast Saccharomyces cerevisiae by Microarray Hybridization", *Molecular Biology of the Cell*, 1998, 9: 3273-3297.

[2] N. Friedman, M. Linial, M., I. Nachman, and D. Pe'er, Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology* ,2000, 7(3): 601-620.

[3] S. Imoto, T. Goto, and S. Miyano, Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression, *Proc Pacific Symposium on Biocomputing*, 2002,7:175-186.

[4] S. Ott, S.Imoto, and S. Miyano, Finding optimal models for small gene networks. *Proc Pacific Symposium on Biocomputing*, 2004, 9:557-567.

[5] P. Brazhnik, A. de la Fuente, and P. Mendes, Gene networks: how to put the function in genomics, *TRENDS in Biotechnology*, 2002, 20(11), 467-472.

[6] H. De Jong, Modelling and Simulations of Genetic Regulatory Network Systems: A Literature Review, *Journal of Computational Biology*, 2002, 9(1): 67-103.

[7] E. Van Someren, L. Wessels, L., and M. Reinders, Genetic network models: A comparative study, *Proc of SPIE*, 2001.

[8] N. Nariai, S. Kim, S. Imoto, and S. Miyano, Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific Symposium on Biocomputing*, 2004, 9: 336-347.

[9] Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S.Miyano, Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 2003, 19(Suppl.2):ii227-ii236.

[10] P.P. Le, A. Bahl, and L.H. Ungar, Using Prior knowledge to improve genetic network reconstruction from microarray data. *In Silico Biolog, 2004, 4(3):335-*

[11] N. Friedman, N., Inferring Cellular Networks using Probabilistic Graphical Models, *Science*, 303 (5659), 2004, 7 99-805.

[12] D. Husmeier, Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian Network, *Bioinformatics*, 2003, 19(17), 2271-2282.

[13] P. H. Lee, and D. Lee, Modularized learning of genetic interaction networks from biological annotations and mRNA expression data, *Bioinformatics*, 2005,21(11), 2739-2747.

[14] S. Kim, S. Imoto, and S. Miyano, Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*. 2004,75(1-3): 57-65.

[15] S. Ott, Finding Optimal Models for gene Networks, PhD Thesis, University of Tokyo, 2004.

[16] N.Friedman, M.Goldszmidt, and A.Wyner, Data Analysis with Bayesian Networks : A Bootstrap Approach, Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence, 1999, pp 196-120.

[17] S.Puri. Small Gene Networks: Finding Optimal Models, Masters Thesis, University of Edinburgh, 2004.